Scientific Research Publishing

Applied Mathematics, 2025, 16(11), 831-850
https://www.scirp.org/journal/am
ISSN Online: 2152-7393
ISSN Print: 2152-7385

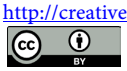
Data Integration Techniques for the Construction of an Integrated Database on the Economic Sustainability of Italian Families

Samuela L'Abbate[1], Paola Perchinunno[2], Antonella Massari[2], Corrado Crocetta[1], Leonardo Salvatore Alaimo[3]

[1]Department of Humanities Research and Innovation, University of Bari Aldo Moro, Bari, Italy
[2]Department of Economics, Management and Business Law, University of Bari Aldo Moro, Bari, Italy
[3]Department of Statistical Sciences, University of Rome La Sapienza, Rome, Italy
Email: samuela.labbate@uniba.it, paola.perchiunno@uniba.it, antonella.massari@uniba.it, corrado.crocetta@uniba.it, leonardo.alaimo@uniroma1.it

How to cite this paper: L'Abbate, S., Perchinunno, P., Massari, A., Crocetta, C. and Alaimo, L.S. (2025) Data Integration Techniques for the Construction of an Integrated Database on the Economic Sustainability of Italian Families. *Applied Mathematics*, **16**, 831-850.
https://doi.org/10.4236/am.2025.1611043

Received: August 2, 2025
Accepted: November 21, 2025
Published: November 24, 2025



Abstract

This work describes a data integration model using the Statistical Matching methodology (hot deck distance) to integrate two surveys conducted by ISTAT (EU-SILC) and the Bank of Italy (Household Income Survey). The construction of an integrated database based on these surveys is useful for studying consumer behavior in relation to specific commodity groups, analyzing household savings decisions, assessing economic and social inequality, and evaluating the impact of public policies through simulations. The coexistence of multiple and diverse objectives necessitates a highly general and versatile integrated file, providing detailed information on spending patterns, savings levels, income distribution, occupational conditions of household members, and more.

Keywords

Statistical Matching, Data Integration Techniques, Record Linkage

1. Introduction

The aim of the work is, therefore, to verify the possibility of building an "integrated database" with information relating to both consumption and income of Italian families, with particular attention to families considered "poor". In this context, "poor" families are understood as those experiencing economic hardship,

typically defined by low-income levels and/or limited consumption capacity, consistent with national poverty thresholds or criteria used by EU-SILC. This definition may include absolute poverty (inability to afford basic goods and services) as well as relative poverty (having significantly less access to resources compared to the national average). In Italy, the main sources of information for the construction of an integrated database of income and consumption are the Eu-Silc survey on Income and Living Conditions (Istat) and the Survey of Household Income and Wealth (SHIW), a long-standing independent microdata source on income and financial behavior conducted by the Bank of Italy.

Since none of the surveys has sufficient coverage to allow the construction of a database of information relating to income and consumption, an attempt was made to integrate the data deriving from the two archives managed by Istat, specifically from EU-SILC and the Bank of Italy's SHIW survey, starting from the assumption that these surveys are more reliable in terms of accuracy of the sample design and control of the representativeness of the sample [1] [2]. It should be noted that although EU-SILC is implemented by Istat in Italy, it is part of a harmonised European statistical program coordinated by Eurostat.

The statistically based methodologies used for the integration of data from multiple sources can be classified into two types: record linkage (exact matching) and statistical matching (statistical matching). Exact matching techniques aim to identify pairs of records related to the same statistical unit, contained in different databases, using specific algorithms [3]; statistical matching techniques, on the other hand, aim to identify records related to similar units, observed in different data archives [4] [5]. Both matching techniques, exact or statistical, allow us to obtain integrated archives, adopting statistically controllable assumptions.

In the case under study, it was decided to cross-reference the records of the families resulting from the Eu-Silc survey and the records of the families resulting from the Bank of Italy survey. It is therefore interesting to verify the possibility of creating integration between the two archives, though always bearing in mind that the available data in both cases concern different samples from the same population: in addition, neither of the surveys covers a sample such as allowing the construction of a single database containing information relating to both income and consumption, a situation well-recognized in international literature as a common limitation of cross-sectional household surveys, particularly in contexts where micro-level budget and income data are independently collected [6].

2. Methods of Integration of Data from Different Sources

2.1. Introduction

The need to use techniques for integrating data from different sources is driven by the increased need for large-scale information and is felt in the most diverse sectors, from official statistics (estimating the size of a population) to epidemiology, where it is commonly used in longitudinal studies to ascertain the effects of certain risk factors, a practice extensively adopted in health services research and

cohort studies [7].

The need to build integrated data archives can be attributed to various reasons [3]:

1) *Occasion:* the creation of complex "welfare state" and taxation systems, which require large databases containing detailed information for individuals and businesses.

2) *Tool:* technology has had a decisive development, making it easier to manage large databases.

3) *Need:* governments have often assumed roles and functions that have increased the need for information, which can be satisfied by the joint use of different sources available, without requiring individuals or businesses to provide information already provided elsewhere. These motivations are also reflected in international statistical guidance, such as the [8] recommendations for combining administrative and survey data sources.

Let us now analyze in detail the two methodologies for integrating data from multiple sources: record linkage and statistical matching which represent two distinct approaches to data fusion, respectively based on unit-level identification and statistical similarity [2] [5].

2.2. Record Linkage

A Record Linkage (R.L.) procedure or exact matching or computer matching is an algorithmic technique whose purpose is to identify which pairs of records, from two databases, correspond to the same statistical unit [4]. It is also defined as an "exact matching technique" because the goal is to exactly connect the units belonging to one database with those of another database [9] a definition consistent with the original probabilistic model introduced by [3].

The main objectives of Record Linkage are:

1) The development of a list of units to be used for the extraction of samples or for the verification of census data (example: the archive of active companies in Italy, ASIA).

2) The reconnection of two or more sources to have a single database, with more information at the unit level.

3) The use of data from various sources as useful information to improve coverage and increase protection against response errors in censuses and surveys.

4) The counting of individuals in a population through capture-recapture methods (for example: the estimate of census under-coverage, carried out by matching census records with those of the post-census survey), a method widely used in demography and epidemiology [10].

The need to connect information relating to the same statistical unit from different sources is increasingly felt in the phases of collection, organization, control and analysis of statistical data. The exact matching of records is a fundamental step for the construction of longitudinal archives and is a preliminary operation for the analysis of the degree of coverage of total surveys and for the control of the

level of errors that may be generated.

Depending on whether the number of records referred to each statistical unit is considered, one-to-one matches can be obtained (each statistical unit corresponds to only one record in each of the archives to be matched), one-to-many (each statistical unit can correspond to multiple records in one of the archives to be matched) or many-to-many (each statistical unit can correspond to multiple records in both archives to be matched).

The main phases identified in a Record Linkage process are:

1. preparation of the input files (pre-processing);
2. selection of common identifying attributes (matching variables);
3. choice of the comparison function;
4. choice of the decision model: estimation of the matching probabilities and evaluation of the model's fit to the data;
5. assignment of the pairs to the "matched" or "unmatched" status;
6. evaluation of the results and selection of unique matches.

The main Record Linkage methods used in different fields of application are the following:

1) Merge by matching: it is based on the ordering of the files to be matched according to a common identification key. It is usually used when the files to be matched belong to the same information system; it is very efficient, although sensitive to errors on the identification key.

2) Deterministic matching: it is based on the concordance of a sufficient number of common variables. For example, if at least two of the three variables (name, surname and year of birth) are concordant, a comparison table is created from which the records that have at least two common variables are chosen. Deterministic matching can consider missing values and errors in the matching variables and allows the informative power of the variables to be graduated using scores established by statistical analysis on external data. The control of possible errors can only be carried out manually (clerical review).

3) Probabilistic matching: as in deterministic matching, we work on the comparison of all possible pairs, using scores based on flexible criteria to establish the matches; The scores and thresholds used to choose the matches depend on the problem at hand. A (optimal) decision rule is established, and the error probability is estimated, this method is formalized in the model by [3] and refined by [11].

Regardless of the technique used, it is always necessary to consider a series of practical problems such as the different data format (standardization of terms), the presence of data duplications (the same unit can be present multiple times in the same file) and the encoding of the variables (parsing), issues widely addressed in the operationalization of record linkage systems, [2].

To use Record Linkage methods, at least two surveys (statistical or administrative) are required that have in common a non-empty set of units and a group of key variables detected in both archives. It is assumed that a subset of k variables, called matching variables, is common to the two archives. In the presence of

matching variables whose combination constitutes a certain and unique key, the problem of accurately recognizing records relating to the same unit is solved by a simple automatic deterministic matching operation. However, often in practice, even when the records in the two archives refer to the same unit, this coincidence does not occur due to missing values or errors in the matching variables. In these cases, it is generally appropriate to use a probabilistic matching procedure [12].

Let's define A and B as two archives made up of v_A and v_B records respectively. We will therefore have:

$$v_a = \{a\}, a = 1, 2, \cdots, v_A$$

$$v_b = \{b\}, b = 1, 2, \cdots, v_B$$

The problem that Record Linkage must solve can be formalized in the following way. Consider all the pairs formed by units from list A and B respectively, if at least one unit is present in both lists; we will have all the pairs formed by units from list A and list B respectively:

$$A \times B = \{(a,b) : a \in A, b \in B\}.$$

The goal is to determine a particular bipartition of the set into two disjoint and exhaustive subsets M and U, such that:

$$M \cap U = \varnothing$$

$$M \cup U = A \times B,$$

where M is the set of pairings of the units detected in A and B, *i.e.* the set of MATCH pairs:

$$M = \{(a,b) : a = b, a \in A, b \in B\}$$

while U is formed by the set of mismatches of the units detected in A and B, or the set of non-pairs (NON MATCH):

$$U = \{(a,b) : a \neq b, a \in A, b \in B\}$$

The fundamental tool used in R.L. is the comparison between the values assumed by the key variables $(X_1, X_2, \cdots, X_k)$ for each pair (a,b) with $a \in A$, $b \in B$.

In general, we represent this comparison with the symbol:

$$y_{a,b} = f\left(x_{a,1}^A, \cdots, x_{a,k}^A; x_{b,1}^B, \cdots, x_{b,k}^B\right)$$

Thus, low levels of diversity are expected for the pairs (a,b) in M and higher levels of diversity for the remaining pairs. The $f(\cdot)$ is an important tool as a discriminant of the quality of the record linkage [5] [13] [14].

It is possible to verify whether a pair (a,b) is a match through the vector of K-component comparisons:

$$y_{ab} = \left(y_{ab}^1, y_{ab}^2, \cdots, y_{ab}^k\right) \text{ con } y_{ab}^h = \begin{cases} 1 \text{ se } x_{a,h}^A = x_{b,h}^B \\ 0 \text{ altrimenti} \end{cases}$$

where the comparison vector takes on the value of one in the case of pairs that

are matches; the value of zero in the case of non-match pairs due to $x_{a,h}^{A} \neq x_{b,h}^{B}$ or in the case in which one of the two elements is missing (partial non-response).

2.3. Statistical Matching

The idea of using statistical matching techniques to combine and enrich information collected in multiple sample surveys is not recent: the first methodological suggestions date back to the early 1960 s. The first significant examples of application are in [15], these early efforts laid the groundwork for subsequent developments in data fusion and synthetic file creation [1].

Statistical matching techniques are used to connect information from two or more data sets, whose units have similarities with respect to a series of variables defined a priori. The underlying assumption is that the sources to be integrated contain both information on a set of common variables and information on a set of distinct variables that are never jointly observed and that it would be interesting to be able to relate. Furthermore, there is never a single result when statistical matching is performed, since much depends on which data set is defined as the basis and on the choices made to "connect" the variables to the basic data set [16] a challenge widely discussed in literature, especially when dealing with non-overlapping variables and sample heterogeneity [5].

An important aspect in statistical matching is therefore represented by the quality of the entire process, which mainly depends on:

1) The quality of the observations coming from the archives.

2) The choice of integration algorithms.

3) Methodologies used to analyze the quantities involved in the integration (correlation/regression coefficients, contingency tables, etc.).

Evaluating quality therefore means understanding how these factors interact and influence the result. The quality of the archive from which the source data comes is fundamental, as it can seriously influence the result. This means that statistical matching can be applied only in those circumstances in which the source data set is characterized by an acceptable level of quality of the information contained, including proper harmonization of variables, consistency of definitions, and handling of missing data [1] [6].

There are two approaches used for Statistical Matching: a micro approach whose objective is to build a complete and synthetic data set and a macro approach whose objective is to estimate the joint distribution of the variables not observed together or of some of its characteristics. The integration of statistical information at the micro level presupposes that the sources from which such information is drawn guarantee a good degree of harmonization, allowing for correct comparability of data at the macro level of aggregates and indicators. A necessary condition is, therefore, that both the units of detection and analysis, as well as the variables that will be used as integration keys, use identical definitions and classifications or those that can be traced back to identities.

In the case of statistical matching the initial assumption is to have *two different*

archives that contain information records on two groups of units selected from the same population. Therefore, the starting point of the issue is the existence of two archives:

1. Containing information on common variables (the socio-demographic type of variables) as well as on different variables that are never been jointly observed,

2. The sources' units to be integrated are separated, or rather the sample surveys are independent among them [17], this is a typical case of "conditional independence" assumption often discussed in the literature [4].

Hence, in the statistical matching the individual records deriving from two or more sources are linked, based on their similarities, through a set of characteristics measured in each source. We consider here two datasets containing two files: file A and file B. To make the statistical matching of these files, it is necessary that common information regarding the units is available in each file. Let X_A be the set of variables measured in file A and let X_B be the set of variables measured in file *B*, it is assumed that these two sets of variables can be transformed in one set with common characteristics. We can indicate the individual's characteristics measured in both datasets as the vector $\boldsymbol{X} = (X_1, \cdots, X_P)$. The remaining variables in each file, that are not overlapping, are indicated as $\boldsymbol{Y} = (Y_1, \cdots, Y_Q)$ in the file *A* and as $\boldsymbol{Z} = (Z_1, \cdots, Z_R)$ in the file *B*.

The common variables *X*, defined from now on as the *integration variables*, are used to identify the units to be linked, while the non-common variables (*Y* e *Z*), defined as the *descriptive variables*, represent the information that is the object of the matching procedures. Therefore, it determines a situation in which the information that is simultaneously gathered on the same units for *Y* and *Z* is missing.

The aim of the statistical matching is to create a file, the *file C* (*the synthetic file*) in which each record contains all variables *X, Y* and *Z*. For each unit in file *A*, a similar unit in file B is identified, whereas the similarity is evaluated in terms of a function of variables *X*. The variables *Z* in file *B* (defined as *the donor*) are then attributed to the matching record in file A, creating thus a *record with full data* (*X, Y, Z*), this is known as the "nearest neighbor donor imputation" strategy [4] [5].

Considering the two archives *A* and *B* (which in this study case are the archives of family consumption and income), we will match the common variables in order to link the descriptive variables.

Some of the methods used for statistical matching are the following:

1) *Hot deck random;*

2) *Hot deck rank;*

3) *Hot deck distance.*

Starting, therefore, from two archives *A* and *B*, we try to create a pairing between the common variables, to connect the descriptive variables. The first method (hot deck random) consists in pairing, in a completely random way, the values relating to the unknown variable *Z* from one archive to another. The fusion of the two archives is obtained by randomly choosing, with repetition, a sample

of n_A units from B and assigning each one to the units of A. By doing so, the possible number of pairings is equal to $n_B^{n_A}$. It is evident that such a solution is justified by a criterion of homogeneity external to the data, which is not entirely satisfactory. A variant of this model can be carried out by restricting the class of donors from B to a defined subset of B, with common variables determining the descriptive variables. For example, let's keep the variable "sex" fixed so that it coincides in both A and B. Let's then randomly select a donor among the units in B that have, for example, the same "sex" as the recipient units in A.

The second method (hot deck rank) consists in connecting the records having values of the distribution function of the common variable that are closest to each other. Specifically, considering X as the common variable, it is necessary to:

calculate the empirical distribution functions of X on A and B:

$$\hat{F}_X^A(x) = \frac{1}{n_A}\sum_{a=1}^{n_A} I(x_a \le x), \quad x \in X$$

$$\hat{F}_X^B(x) = \frac{1}{n_B}\sum_{b=1}^{n_B} I(x_b \le x), \quad x \in X$$

for each $a = 1, 2, \cdots, n_A$, the record b^* is associated in B such that:

$$\left|\hat{F}_X^A\left(x_a^A\right) - \hat{F}_X^B\left(x_{b^*}^B\right)\right| = \inf_b \left|\hat{F}_X^A\left(x_a^A\right) - \hat{F}_X^B\left(x_b^B\right)\right|$$

The method applied in this work is the *hot deck distance* which consists of matching the record a from A with the record b^* from B that is the "closest" considering the variety of common variables. Hence, we will have that:

$$d_{a,b^*} = \left|x_a^A - x_{b^*}^B\right| = \min_{1\le b\le n_B}\left|x_a^A - x_b^B\right|$$

Each unit of the base dataset (*i.e.* archive defined as the receiving one) is operatively associated to another unit of the second dataset (*i.e.* the archive defined as the donor) by applying the *distance function* $d(x_i, x_j)$ which is computed on the integration variables and assumes much lower values as the individuals are more similar among themselves (*nearest neighbor match*), a common approach in applied social science and economic datasets [6].

In the case of qualitative or categorical variables, distance measurement can be done in different ways. In general, when dealing with variables of this nature, it is customary to think in terms of similarity (similarity) between units rather than distance.

In the presence of binary categorical variables (presence/absence of a certain characteristic), well-known similarity measures are [18]:

1. Matching coefficient:

$$S_M(A,B) = \frac{p_{00} + p_{11}}{p}.$$

where p_{00} and p_{11} are respectively the number of variables for which A and B jointly present modality 0 and modality 1; this coefficient is the fraction of the p variables that present the same modality.

2. Jaccard coefficient:

$$S_J(A,B) = \frac{p_{11}}{(p - p_{00})}$$

An extension of the matching coefficient for non-binary categorical variables is the following:

$$S_M(A,B) = \frac{n(x_A \cap x_B)}{p}.$$

where $n(x_A \cap x_B)$ indicates the number of characteristics common to the two units.

Note that the number of different characteristics between the two units is one of the simplest distance measures; it is also called Hamming distance:

$$d_H(A,B) = p - n(x_A \cap x_B) = \sum_{k=1}^{p} \delta(x_{Ak}, x_{Bk})$$

with a value equal to zero if A and B present the same modalities and a value equal to one otherwise:

$$\delta(x_{Ak}, x_{Bk}) = \begin{cases} 0, & se \quad x_{Ak} = x_{Bk} \\ 1, & se \quad x_{Ak} \neq x_{Bk} \end{cases}$$

In literature (especially in the field of cluster analysis or discriminant analysis) in some cases it is suggested to transform the categorical variables into quantitative ones, to be able to apply the Minkowsky metric or the Mahalanobis distance.

Each pair identified will give rise to an integrated record, containing information collected in both surveys. The resulting integrated file will be representative of the same population as the base file but will also have observations on new variables coming from units similar to those present in the base file [17].

3. The Integration of the EU-SILC and Bank of Italy Archives

3.1. Introduction

The reasons underpinning the development of an integrated database from the two surveys under study are numerous. In particular, the possible uses of the integrated file include the study of consumer behavior in relation to specific groups of goods, the analysis of decision-making on household savings, the analysis of economic and social inequality as well as the study of the impact of public policies through simulations. The coexistence of multiple and diverse objectives creates the need to obtain a highly generalized and versatile integrated file, which provides detailed information each time on the various items of expenditure, savings rates, income distribution, employment conditions of family members and so on, a flexibility increasingly demanded in microsimulation models for social and tax policy analysis.

Since the two surveys confirm a significant difference between the samples, it is necessary to identify a file of a survey as a base (typically that considered more reliable from the point of view of their representativeness) using another file as a

donor. The ISTAT survey proves to be more reliable than that of the Bank of Italy in terms of accuracy of the sample design and control of the representativeness of the sample yet matching each ISTAT unit with a similar Bank of Italy unit would require replicating each Bank of Italy unit three times with obvious consequences in terms of variability in the integrated file. Accepting this level of replication would have significantly increased the sampling variability and compromised the stability of derived indicators, making the ISTAT file a less desirable base despite its higher representativeness. The choice of the Bank of Italy as a base file would not, however, present such problems, an issue also linked to the relative sample sizes and survey design effects, which may amplify the variance introduced by duplication [1]. For these reasons, an alternative matching may involve matching each Bank of Italy unit with a selection of spending variables from the ISTAT file. These variables should not be those originally reported (excessively detailed) but rather, their aggregation into appropriate consumption categories, similar to those of the standard files, a step aligned with international practice for harmonizing consumption data across surveys.

Among the methods, the most appropriate method for the integration between the two files under study appears to be that based on the definition of a distance function (distance hot deck), which allows for checking the similarity between the units of the first and second sample by assigning to each unit of a sample a sufficiently similar unit of the other sample. The unit to be matched will not necessarily be at zero distance from the unit under observation, but the threshold values for the distance function may be determined. Each pair thus identified will create an integrated record, containing both common information, as detected in both surveys, as well as those specific to each source, a matching logic widely implemented using nearest-neighbor or propensity score-based techniques in synthetic data generation [5] [6].

3.2. Harmonization of Common Information

The first phase necessary for carrying out of statistical matching is the so-called stage of harmonization of information deriving from the two archives. This process is essential to check the real level of comparability in the survey. The individual steps of the process of harmonization may be summarized as:

1. *Analysis of input sources.* Input sources arising from individual archives are analyzed in terms of the quality of key information gathered, completeness and eligibility.

2. *Selection of variables from input sources.* Two datasets deriving from different archives are prepared by selecting only common variables (e.g. sex, age, marital status, etc.) and the integration variables (such as income and consumption).

3. *Harmonization of common variables.* The common variables selected are aligned in terms of classification. For example, in the Bank of Italy survey, the code relating to educational qualifications is identified with codes increasing according to the increasing level of education (no qualification, primary school, sec-

ondary school, vocational diploma, high-school diploma, university diploma, degree and post-graduate studies), while the codes in the ISTAT survey decrease. This leads, therefore, to the reversal of the codes of one of the two sources. This step is widely documented in statistical data integration literature as "semantic harmonization", and it is considered essential to ensure structural comparability across datasets [5] [14] [19].

Compared to the random and rank-based hot deck methods, the distance hot deck approach was preferred as it ensures a higher degree of similarity between matched units across multiple integration variables. This feature is particularly aligned with the goals of this study, which require preserving the joint structure of socio-demographic characteristics when imputing income and consumption data, a critical factor for reliable economic and inequality analyses.

Starting, therefore, from the two datasets of harmonized information, the following choices were made:

1. Identification of the *stratification variables;*
2. Identification of *integration variables* (or matching);
3. Identification of *descriptive variables* (or analysis).

Specifically, given:

1) Set1 = "*Bank of Italy*".

2) Set2 = "EU-SILC".

The *stratification variables* were defined as:

1) Region of residence of the family;

2) Number of members per household.

With the *integration variables* (common variables) as the following:

1) Matching (1) = gender.

2) Matching (2) = age classes.

3) Matching (3) = marital status.

4) Matching (4) = educational qualification.

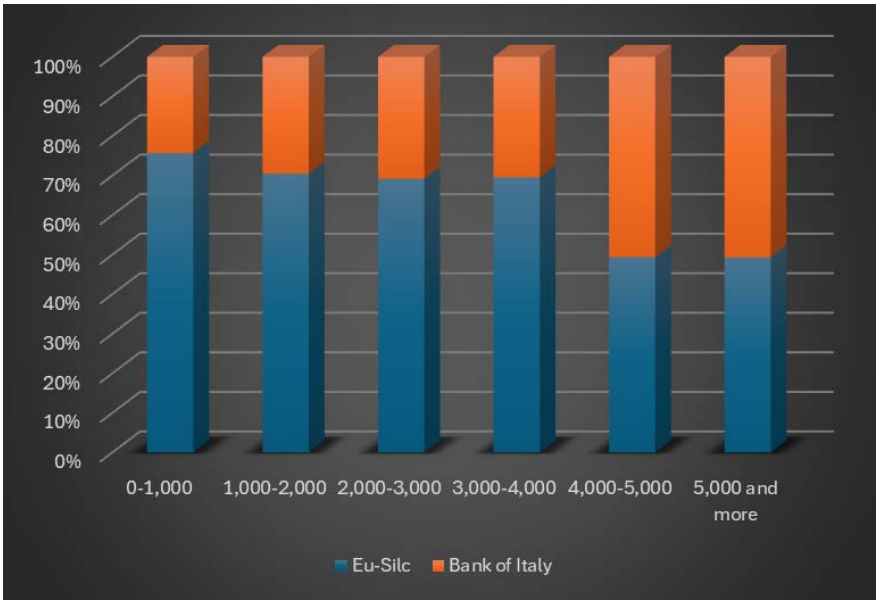
Finally, the different information obtained from the two different datasets was matched. In particular, the following descriptive variables from the EU-SILC survey archive were considered: total income, income from employment, income from self-employment and retirement income. Information relative to the following descriptive variables was, however, considered from the Bank of Italy archive: income from employment, income from self-employment, income from pensions and transfers, capital income and savings capacity. Such variable selection reflects standard practices in statistical matching applications where variables are grouped into "matching variables" and "non-overlapping variables" to be imputed across files [1] [5].

As previously stated, a threshold value was defined for a highly restrictive distance function, equal to 0 (specifically, considering the 4 matching variables only the matches with a distance equal to 0 were considered) in order to consider all the matching variables as valid. This corresponds to exact matching under deterministic distance, often referred to in literature as "constrained hot deck" with strict similarity threshold [6].

3.3. First Results of the Integration of Archives

Having completed the matching procedure, evaluation was carried out of validity in terms of both the preservation of the distribution between the different cells identified by vector $\boldsymbol{X}$, and in terms of maintaining pre-existing relationships between the variables of interest. This step, often referred to as "consistency checking" or "distribution preservation validation", is essential in statistical matching and typically involves the comparison of marginal and joint distributions across datasets [5] [14]. A comparison was then made between the percentage distribution relative to the region of residence and the number of family members in the integrated dataset and in those of the EU-SILC and Bank of Italy surveys.

The integrated archive is, therefore, composed of several records distributed homogeneously in comparison to those of origin. The comparison between the descriptive variables concerning the income component is of particular significance. Proceeding to classification of the total household income into classes it may be noted that the data from the two base archives differ from each other in terms of income distribution (Figure 1).



Source: Our elaboration of integrated archive data of EU-SILC and Bank of Italy.

Figure 1. Composition percentage of records of EU-SILC and the Bank of Italy surveys by family income class.

Furthermore, small differences emerge from a comparison of the distribution of incomes of the two base archives with the integrated archive (Table 1). In particular, the integrated archive presents, in comparison to EU-SILC, an underestimation of low-income households (7.5% compared to 13.4% for EU-SILC) and an overestimation of families with incomes of between €3,000 and €4,000 (19.9% against 14.7% for EU-SILC). Conversely, the integrated archive, in comparison to the Bank of Italy archive, presents an overestimation of low-income households

(12.7% against 10.4% for the Bank of Italy) and an underestimation of households with incomes of between €2,000 and €3,000 euro (22.9% compared to 25% for the Bank of Italy). Such discrepancies are commonly encountered in statistical matching and are often attributed to the non-identical sampling designs and response behaviors across source surveys [1].

Table 1. Composition percentage of records of EU-SILC and the Bank of Italy surveys and integrated archives by family income class.

Income Classes	EU-SILC	Integrated archive	Bank of Italy	Integrated archive
0 - 1000	13.4	7.5	10.4	12.7
1000 - 2000	31.1	26.4	31.4	31.9
2000 - 3000	23.3	26.1	25.0	22.9
3000 - 4000	14.7	19.9	15.5	14.4
4000 - 5000	8.4	9.7	8.2	8.6
Oltre 5000	9.1	10.4	9.5	9.4
Total	**100**	**100**	**100**	**100**

Source: Our elaboration of integrated archive data of EU-SILC and Bank of Italy.

These differences, particularly the underestimation of low-income households in the integrated file compared to the EU-SILC archive, may have relevant implications for the subsequent cluster analysis. Specifically, poverty profiles identified in Section 4 may be biased toward moderate, or higher, income groups, potentially reducing the visibility of the most vulnerable segments of the population. As a result, policy recommendations drawn from the clusters should be interpreted with caution, especially when targeting interventions toward households in extreme economic hardship.

4. The Construction of Hardship Profiles

4.1. Cluster Analysis

The next step of the integrated archive analysis was to consider a clustering procedure with the objective of outlining various poverty profiles, not defined a priori, to assign each family with socio-economic behavior resulting from the matching between the two archives [20].

Cluster analysis is highly effective since it provides "*relatively distinct*" (or heterogeneous) clusters, each consisting of units (families) with a high degree of "*natural association*", a property that makes clustering particularly suitable for socio-economic segmentation [21].

The different approaches to cluster analysis share a common need to define a matrix of dissimilarity or distance between the n pairs of observations, which represents the point at which each algorithm is generated [17] [22].

The most recent studies in the field of data mining are directed towards the search for algorithms able to deal with both very large datasets and datasets consisting of mixed variables. An algorithm of this type is the k-prototypes [23] [24] in which it is assumed that the measure of dissimilarity on numerical attributes is defined by Euclidean distance, while that of the categorical attributes s_c is defined as the number of "wrong unions of categories" between two objects. The measure of dissimilarity between two objects is, therefore, defined as:

$$\gamma s_n + (1-\gamma) s_c$$

where, γ is a weight that has the objective of avoiding a favoring of one or the other type of attribute. One downside of such an algorithm lies precisely in the arbitrariness in the choice of a suitable weighting.

Several extensions to the k-prototypes model have been proposed to handle uncertainty in mixed data, such as fuzzy clustering approaches or automated λ selection [25].

A cluster analysis technique that shows no signs of weakness in this sense is defined as *TwoStep*. This is an extension of the distance measures used by [26] based on the model introduced for data with continuous attributes.

The *TwoStep* algorithm has two advantages: it treats variables of mixed type and automatically determines the optimal number of clusters, although it allows for fixing the desired number of clusters [27].

The *TwoStep* procedure, highly efficient for large datasets, is a cluster analysis scalar algorithm and is able to simultaneously treat variables or categorical and continuous attributes. This is achieved through two steps:

1. In the *first step*, defined as pre-cluster, records are pre-classified into many small sub-clusters;

2. In the *second step* the sub-clusters (generated in the first step) are grouped into several clusters that optimize the BIC (*Bayesion Information Criterion*) [28] [29].

The *pre-clustering* phase is a process of segmentation in which the results of the algorithm may result in an initial partition of the space where the variables are defined (considering the order of their importance) or the distance between the cases. This partition is represented by a tree known as the *Cluster Features Tree* defined by levels of nodes. All cases, starting from the node to the root, are channeled through other nodes until they become terminal nodes as they consist of very close cases (within a distance threshold). If there is a suitable match, the record most "distant" from the others is used to begin a terminal node itself. If the terminal node exceeds the distance threshold, the terminal node is split into two, using the farthest pair according to the selected criterion and redistributing the remaining ones on the basis of the criterion of proximity. If this recursive process helps the tree to grow beyond the full extent granted or, rather, within the calculation memory limits employed, it is reconstructed based on the existing one, increasing the distance threshold, thus allowing the entry of new records. This process ceases when all records have been examined.

In the *second step* the sub-clusters produced in the pre-clusters are further classified. In this second stage, given the modest dimensions, traditional methods of clustering can also prove effective.

The *TwoStep* considers the optimal partition by using the *Bayesian Information Criterion* (BIC) that for k cluster is defined as:

$$BIC_K = -2l_k + r_k \log n$$

where r_k is the number of independent parameters and $l_k = \sum_{v=1}^{k} \xi_v$ is the function of log-likelihood, for the step with k clusters, which can be interpreted as the is the function of log-likelihood, for the step with k clusters, which can be interpreted as the dispersion within the clusters. It also represents the entropy within the k clusters in the case in which only categorical variables are considered (ξ_v will be analyzed below) [30] [31].

4.2. Application with a Clustering Technique for Mixed Variables

The clustering technique has allowed for outlining various poverty profiles with which to associate the families of the EU-SILC-Bank of Italy integrated archive.

In particular, the characteristics of the different clusters relate to the composition of the family, their educational level, the age of the respondent and the age group, detected in the two surveys. This therefore leads to the following profiles:

1) *Cluster 1: Pensioners.* The main profile is characterized by households with one or two components (widows or widowers), with a head of the family with a low level of education and more than 61 years of age, whose family income is less than € 2,000 per month. Typical profile of elderly individuals with limited pensions, at risk of economic hardship.

2) *Cluster 2: Large families.* The main profile is characterized by large families, with a head of the family with a low level of education and more than 46 years of age, whose family incomes are differentiated, but greater than €2,000 per month. Economic risk due to expenses for children and family management.

3) *Cluster 3: Graduates.* The main profile is characterized by single family members or small families with a head of the family with a high educational level and more than 31 years of age, whose incomes are differentiated, and are, in many cases greater than €5,000 Euros per month. High socio-economic status profile, with good financial stability.

4) *Cluster 4: Low level of education*. The main profile is characterized by single family members or small families (60% with less than two members), with a head of the family with a low level of education and more than 46 years of age, whose incomes are differentiated. Vulnerable profile, facing economic difficulties due to low education and limited job opportunities.

5) *Cluster 5: High-school graduates.* The main profile is characterized by small families, with a head of the family with high-school level education and more than 31 years of age, whose incomes are higher than those in cluster 4 (50% with incomes of over €3,000 per month). Intermediate situation between financial stability and economic vulnerability.

Cluster Analysis allowed for the identification of distinct groups of families with similar characteristics in terms of income, education, and family structure. The results highlight how the economic situation of Italian families is strongly influenced by factors such as level of education, number of household members and age of the head of household, see also [32] [33], who underline the multidimensionality of poverty profiles in EU and OECD contexts).

The clusters also reveal potential economic hardship situations, such as retirees with low pensions and large families with high financial burdens. In contrast, families where the head has a university degree tend to have more stable financial conditions (consistent with findings in [34] regarding income resilience and human capital).

This analysis can be useful in guiding social and fiscal policies aimed at reducing economic inequalities and supporting the most vulnerable categories [35].

5. Conclusions

This paper aimed to construct an integrated database combining income and consumption data from two major Italian surveys—EU-SILC and the Bank of Italy's SHIW—using statistical matching techniques. By employing the hot deck distance method, we generated a synthetic dataset that allows for richer analyses of household economic conditions, particularly in identifying profiles of economic hardship. This has led to an in-depth analysis of the phenomenon, from both an IT and statistical perspective, see [2] [5] for a comprehensive review of matching techniques in official statistics.

The reasons for constructing an integrated database from the two surveys analyzed are numerous. These include analyses of household saving behavior, studies on economic and social inequalities, and assessments of public policy impacts through data simulations. The presence of multiple differentiated objectives necessitates the creation of a flexible and detailed integrated file, providing useful information on expenditures, savings rates, income distribution, and employment conditions of family members. The availability of high-quality individual data on income and consumption is, therefore, more essential than ever [8].

Statistical Matching is a valuable tool for integrating data from multiple sources [13]. However, in analyzing this methodology, some critical issues emerge, including:

1) Harmonization of data sources, a process that is costly in terms of both time and economic resources, especially when sources use different measurement scales for common variables or differing approaches in handling missing data and error correction [14] [19].

2) Choice of integration model, which varies depending on the type of available information. If auxiliary information is available, it should be used in the Statistical Matching process, as it provides fundamental knowledge of the studied phenomenon [1] [36].

3) Evaluation of the quality of the integrated database, which remains an open

challenge. Although methods exist to estimate standard deviation and variance, there is still room for improvement, particularly in estimating variance resulting from imputing missing data [37].

Creating an integrated dataset is not only beneficial for end users but also for data producers and managers. Methodological comparisons, analysis of informational needs, and the complementarity of different surveys can serve as important stimuli for improving the quality of statistical surveys [38].

The results obtained demonstrate how it is possible to identify social stratifications based on different factors that simultaneously influence consumption behavior and average family income, such as household composition, territorial area of residence, and the occupational status of the household head. In particular, the simulation results allow for the identification of well-defined profiles of economic hardship in terms of family type and territorial location [32] [35].

Through this analysis, an attempt was made to quantify the influence of geographical distribution and household type (number of family members) on income determination. While estimates of poverty risk based on objective indicators such as income or consumption expenditure are essential, it is equally useful to detect families' subjective perception of their standard of living and economic difficulties. These subjective factors provide valuable insights not only for better illustrating and understanding the poverty phenomenon but also for informing public policies aimed at combating it [39] [40].

In recent years, both in scientific and political fields, there has been growing interest in poverty and, more generally, in social exclusion phenomena. However, many studies face a lack of detailed statistical data especially in relation to multidimensional poverty measures, as discussed in [41].

Poverty analysis is often based on macro-level general surveys. However, in Italy, conducting micro-level surveys, as has been done for years in the United States, would be of particular interest. This would allow for the refinement of social policies and the development of more effective strategies to combat inequality [33] [42].

It is hoped that the differences in poverty profiles emerging from analyses conducted using different criteria may serve as a stimulus for developing new solutions, capable of offering benefits not only to the scientific community but to society.

Despite some limitations, such as the underestimation of low-income households due to sample differences, the integrated archive provides a valuable tool for socioeconomic research and the design of targeted public policies. Future work should further refine matching techniques and incorporate additional sources to enhance the precision of poverty measurement and policy simulations.

Acknowledgments

This study was funded by the European Union-NextGenerationEU, Mission 4, Component 2, in the framework of the GRINS-Growing Resilient, Inclusive and

Sustainable project (GRINS PE00000018-CUP H93C22000650001). The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union, nor can the European Union be held responsible for them.

The authors would also like to thank the Italian National Institute of Statistics (ISTAT) and the Bank of Italy for providing access to the data used in this research.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Rässler, S. (2002) Statistical Matching: A Frequentist Theory, Applied Bayesian Methodology and a New Approach. Springer, 266 p.

[2] Christen, P. (2012) Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer, 221 p.

[3] Fellegi, I. (1997) Record Linkage and Public Policy: A Dynamic Evolution. In: Alvey, W. and Jamerson, B. Eds., *Record Linkage Techniques*, Arlington, 3-12.

[4] Belin, T.R. and Rubin, D.B. (1995) A Method for Calibrating False-Match Rates in Record Linkage. *Journal of the American Statistical Association*, **90**, 137-147. https://doi.org/10.2307/2291082

[5] D'Orazio, M., Di Zio, M. and Scanu, M. (2002) Statistical Matching and Official Statistics. Quaderni di Ricerca ISTAT, 1.

[6] Conti, P. and Marella, D. (2014) Uncertainty in Statistical Matching for Complex Sample Surveys. 47*th Meeting of the Italian Statistical Society*, Cagliari, 11-13 June 2014, 1-6.

[7] Bohensky, M.A., Jolley, D., Sundararajan, V., Evans, S., Pilcher, D.V., Scott, I., *et al.* (2010) Data Linkage: A Powerful Research Tool with Potential Problems. *BMC Health Services Research*, **10**, Article No. 346. https://doi.org/10.1186/1472-6963-10-346

[8] UNECE (2021) Guidelines on Statistical Data Integration. United Nations Economic Commission for Europe, 174 p. https://unece.org/info/publications/pub/36715

[9] Schionato, L. (1995) Tecniche di linkage statistico per il raccordo di una pluralità di fonti amministrative. In: Biffignandi, S. and Maritni, M., Eds., *Il Registro Statistico Europeo Delle Imprese*, Franco Angeli Editore, 333-343.

[10] Winkler, W.E., Yancey, W.E. and Porter, E.H. (2006) The Optimizer's Curse: Skepticism and Postdecision Surprise in Decision Analysis. In *Advances in Record-Linkage Methodology as Applied to Matching the* 1985 *Census of Tampa*, *Florida*. U.S. Census Bureau.

[11] Jaro, M.A. (1989) Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, **84**, 414-420. https://doi.org/10.2307/2289924

[12] Nuccitelli, A., Bosio, F. and Fioriti, L. (2004) L'applicazione reclink per il record linkage: Metodologia implementata e linee guida per la sua utilizzazione. Istat.IT.

[13] D'Orazio, M., Tzavidis, N. and Salvati, N. (2021) Framework for Statistical Matching with Auxiliary Information. *Statistical Methods & Applications*, **30**, 861-885.

[14] Fortini, M., Liseo, B. and Scanu, M. (2002) On Bayesian Record Linkage. *Research in Official Statistics*, **5**, 185-198.

[15] Ruggles, N.N. and Ruggles, R. (1974) A strategy for Merging and Matching Micro Data Sets. *Annals of Economic and Social Measurement*, **3**, 353-371.

[16] Fortunato, E. and Morrone, A. (1999) Approcci micro e macro al linkage dei dati delle indagini ISTAT sulle famiglie. Atti del convegno SIS: Verso i censimenti del 2000, 253-260.

[17] Montrone, A., Perchinunno, P. and de Blasi, R. (2012) Statistical Matching of EU-SILC and HBS Data. *Statistica & Applicazioni*, **10**, 113-130.

[18] Ryu, T. and Eick, C. (1998) A Graph-Based Approach for Discovering Various Kinds of Association Rules. *Proceedings of the* 1998 *ACM Symposium on Applied Computing*, Atlanta, 27 February-1 March 1998, 260-267.

[19] Eurostat (2017) Handbook on Data Integration Methods. Eurostat Manuals and Guidelines, 158 p.

[20] Zhang, T., Ramakrishnan, R. and Livny, M. (1996) BIRCH: An Efficient Data Clustering Method for Very Large Databases. *ACM SIGMOD Record*, **25**, 103-114. https://doi.org/10.1145/235968.233324

[21] Everitt, B.S., Landau, S., Leese, M. and Stahl, D. (2011) Cluster Analysis. Wiley. https://doi.org/10.1002/9780470977811

[22] Chiu, T., Fang, D., Chen, J., Wang, Y. and Jeris, C. (2001) A Robust and Scalable Clustering Algorithm for Mixed Type Attributes in Large Database Environment. *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 26-29 August 2001, 263-268. https://doi.org/10.1145/502512.502549

[23] Huang, Z. (1997) A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining. *Data Mining and Knowledge Discovery*, **3**, 34-39.

[24] Huang, Z. (1998) Extensions to the K-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery*, **2**, 283-304. https://doi.org/10.1023/a:1009769707641

[25] Vathy-Fogarassy, Á. and Abonyi, J. (2013) Graph-Based Clustering and Data Visualization Algorithms. Springer.

[26] Banfield, J.D. and Raftery, A.E. (1993) Model-Based Gaussian and Non-Gaussian Clustering. *Biometrics*, **49**, 803-821. https://doi.org/10.2307/2532201

[27] Vermunt, J.K. and Magidson, J. (2002) Latent Class Cluster Analysis. In: Ha-Genaars, J.A. and Mcphee, I., Eds., *Applied Latent Class Analysis*, Cambridge University Press, 89-106. https://doi.org/10.1017/cbo9780511499531.004

[28] Schwarz, G. (1978) Estimating the Dimension of a Model. *The Annals of Statistics*, **6**, 461-464. https://doi.org/10.1214/aos/1176344136

[29] Fraley, C. and Raftery, A.E. (2002) Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association*, **97**, 611-631. https://doi.org/10.1198/016214502760047131

[30] Pelleg, D. and Moore, A.W. (2000) X-Means: Extending K-Means with Efficient Estimation of the Number of Clusters. *Proceedings of the Seventeenth International Conference on Machine Learning* (*ICML* 2000), Stanford, 29 June 2000-2 July 2000, 727-734.

[31] Steinley, D. (2006) K-Means Clustering: A Half-Century Synthesis. *British Journal of Mathematical and Statistical Psychology*, **59**, 1-34. https://doi.org/10.1348/000711005x48266

[32] Whelan, C.T., Layte, R. and Maître, B. (2003) Persistent Income Poverty and Deprivation in the European Union: An Analysis of the First Three Waves of the European Community Household Panel. *Journal of Social Policy*, **32**, 1-18. https://doi.org/10.1017/s0047279402006864

[33] Tanton, R., Vidyattama, Y., Mcnamara, J. and Vu, Q.N. (2009) Small Area Estimation for Local Economic Indicators in Australia. *Australasian Journal of Regional Studies*, **15**, 303-325.

[34] Jenkins, S.P. and van Kerm, P. (2009) The Measurement of Economic Inequality. In: Salverda, W., Nolan, B. and Smeeding, T., Eds., *The Oxford Handbook of Economic Inequality*, Oxford University Press, 40-67.

[35] Nolan, B. and Whelan, C.T. (2011) Poverty and Deprivation in Europe. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199588435.001.0001

[36] D'Orazio, M., Di Zio, M. and Scanu, M. (2006) Statistical Matching. Wiley. https://doi.org/10.1002/0470023554

[37] Little, R. and Rubin, D. (2019) Statistical Analysis with Missing Data. 3rd Edition, Wiley. https://doi.org/10.1002/9781119482260

[38] OECD (2022) Enhancing the Use of Administrative Data in Official Statistics. OECD Statistics Working Papers, No. 3.

[39] OECD (2019) Measuring the Effectiveness of Social Protection. Social Policy Report, 108.

[40] EUROFOUND (2020) Addressing Household Over-Indebtedness. Publications Office of the European Union, 92 p.

[41] Alkire, *et al.* (2015) Multidimensional Poverty Measurement and Analysis Get Access Arrow. Oxford University Press.

[42] Atkinson, A.B. (2019) Measuring Poverty around the World. Princeton University Press, 464. https://doi.org/10.2307/j.ctvc77fd6